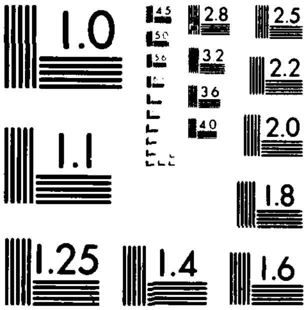
AD-A114 573 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/6 12/1
AN EMPIRICAL BAYESIAN APPROACH TO THE SMOOTH ESTIMATION OF UNKN--ETC(U)
FEB 82 T LEONARD DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2339 NL

END
DATE
FILMED
6 82
DTIC


1.0
4.5
5.0
5.6
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

2

MRC Technical Summary Report #2339

AN EMPIRICAL BAYESIAN APPROACH TO THE SMOOTH ESTIMATION OF UNKNOWN FUNCTIONS

Tom Leonard

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

February 1982

Received January 11, 1982

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

AN EMPIRICAL BAYESIAN APPROACH TO THE SMOOTH ESTIMATION OF UNKNOWN FUNCTIONS

Tom Leonard

Technical Summary Report #2339
February 1982

ABSTRACT

A Bayesian procedure is described for smoothly estimating unknown functions, given a finite set of observations. It is assumed that a suitable transformation of the function can be taken to possess a Gaussian prior distribution across function space. The five special cases (a) estimation of a logistic density transform, (b) the log intensity function of a non-homogeneous Poisson process, (c) the log hazard function for survival data, (d) the logit function in biossay, and (e) the mean value function in a possibly non-linear time series of the Kalman type or equivalently a regression function for possibly non-normal observations, are considered, and in each case a non-linear Fredholm equation is described for the posterior estimate. In cases (d) and (e) this reduces to a finite dimensional system. In all five cases an approximate procedure is developed which is particularly useful when the sample size is large. This approximates the function space prior by a multivariate normal prior on the coefficients in a linear approximation, and then proceeds by conventional Bayesian techniques. In cases where the prior covariance kernel is assumed to posses a particular parametrized form (e.g. from the Orstein-Ulenbeck process) the approximations enable us to estimate the prior parameters appearing in the kernel, empirically from the data, via a lemma relating to the EM algorithm. Finally, in a very special case, involving normal observations and an integrated Wiener prior, some fresh Bayesian and empirical Bayesian results are developed.

AMS(MOS) Subject Classifications: 62G05, 62M10, 62H12

Key Words: Empirical Bayes; function space; density estimation; non-homogeneous Poisson process; survival data; hazard function; biossay; logit; time series; non-linear filtering; Kalman; Gaussian process; Orstein-Uhlenbeck; covariance kernel; EM algorithm.

Work Unit No. 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Prior function space/non-linear Fredholm equation procedures are considered for the smooth estimation of unknown functions given finite sets of observations. Situations considered include (a) density estimation, (b) estimation of the intensity function for a non-homogeneous Poisson process, (c) the hazard function for survival data, (d) the logit function in biossary, and (e) the regression function or time-varying mean value function for possibly non-normal observations. The problem may be approximately reduced to a Bayesian analysis in finite-dimensional space thus leading to simple approximate solutions to the Fredholm equations. This facilitates the empirical estimation of the smoothing and shrinkage parameters appearing in the prior covariance kernel, via a lemma relating to the EM algorithm. Finally, in a well-known special case, the estimation of a normal mean value function under an integrated Wiener prior is considered and some fresh procedures are described for making inferences about the smoothing parameter λ.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

AN EMPIRICAL BAYESIAN APPROACH TO THE SMOOTH ESTIMATION OF UNKNOWN FUNCTIONS

Tom Leonard

1. INTRODUCTION.

We consider situations where the likelihood, given the observation vector $\underset{\sim}{y}$, acts as an integral operator upon an unknown function $\lambda(t)$, for $t \in (a,b)$. In many such situations it is possible to find a mapping ξ such that

$$\lambda = \xi(g)$$

where $g(\cdot):(a,b) \to R^1$ is not subject to any functional or inequality constraints. The problem considered in this paper is the smooth estimation of g and hence λ, given the observation vector $\underset{\sim}{y}$.

Our approach will, for example, be applicable to the following situations:

(a) Let $\underset{\sim}{y} = (y_1,\cdots,y_n)^T$ where $y_1,\cdots,y_n$ constitute a random sample from a distribution with unknown density $\lambda(t)$ for $t \in (a,b)$. Then (see Leonard, 1978) it is often useful to work in terms of a logistic density transform g satisfying

$$\lambda(t) = e^{g(t)}/\int_a^b e^{g(s)}ds \tag{1.2}$$

which avoids the constraints that λ is non-negative and integrates to unity. In this case the log-likelihood functional of g, given $\underset{\sim}{y}$ is denoted by

$$L(g) = \sum_{i=1}^{n} g(y_i) - n \log \int_a^b e^{g(s)}ds.$$

(b) Let $y_1,\cdots,y_n$ denote the arrivals during a fixed interval (a,b) for a time-varying Poisson process with unknown rate function $\lambda(t)$ for $t \in (a,b)$.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

Here n is itself a random variable and, for $n > 1$, the log-likelihood of $g = \log \lambda$, given $\underset{\sim}{y}$ and n, is

$$L(g) = \sum_{i=1}^{n} g(y_i) - \int_a^b e^{g(s)}ds. \qquad (1.3)$$

(c) Let $\underset{\sim}{y} = (y_1,\cdots,y_r;y_{r+1},\cdots,y_n)^T$, where r is a random variable, with $y_{r+1},\cdots,y_n$ denoting fixed censoring times, and suppose we are concerned with the estimation of the survivor density $\lambda(t)$ for $t \in (0,\infty)$, from which the uncensored observations $y_1,\cdots,y_r$ are a random sample. Let

$$h(t) = \lambda(t)/(1 - \Lambda(t))$$

with

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

denote the hazard function. Then the log-likelihood of the log-hazard function $g(t) = \log \Lambda(t)$ is given by

$$L(g) = \sum_{i=1}^{r} g(y_i) - \sum_{i=1}^{n} \int_0^{y_i} e^{g(s)}ds. \qquad (1.4)$$

(d) Consider the classical biossay problem where n rats receive dose levels $x_1,\cdots,x_n$ and we wish to estimate the function $\lambda(t)$ for $t \in (0,\infty)$ which represents the probability that a rat will die if it receives dose level t. Let

$$y_i = \begin{cases} 1 & \text{if rat with dose level } i \text{ dies} \\ 0 & \text{otherwise.} \qquad (i = 1,\cdots,n) \end{cases}$$

Then the log-likelihood of the logit function

$$g(t) = \log \theta(t) - \log[1 - \theta(t)]$$

is given by

$$L(g) = \sum_{i=1}^{n} y_i g(x_i) - \sum_{i=1}^{n} \log\{1 + e^{g(x_i)}\}. \qquad (1.5)$$

(e) Suppose that we observe n independent observations $y_1,\cdots,y_n$ relating to an unconstrained function $g(t)$ for $t \in (a,b)$ such that the density or probability mass function $b(y_i;g(t_i))$ of y_i given g, depends upon $g(t_i)$ but not further upon g.

For example, t might denote a time variable and the y_i might be independent and Poisson distributed with mean equal to $\exp\{g(t_i)\}$. When we add a further distribution to g we see that this will yield a time series model for Poisson observations, with possibly unequal time points. Alternatively $t_1,\cdots,t_n$ might correspond to explanatory variables so that we are in a non-linear regression situation. In any case

$$L(g) = \sum_{i=1}^{n} \log b(y_i;g(t_i)) \qquad (1.6)$$

for which (1.5) is a special case in the binary situation.

In all five cases it is necessary to assume some prior information in order to estimate the whole curve g, given only a finite set of observations. A wide range of prior distributions are contained in the Gaussian family. In each case it might be reasonable to take $\{g(t);t \in (a,b)\}$ to assume the probability structure of a Gaussian process with mean value function $\mu(t)$, and covariance kernel

$$\mathrm{cov}(g(s),g(t)) = K(s,t) \quad (s \in (a,b);\ t \in (a,b)). \qquad (1.7)$$

For example, the choice

$$\mathrm{cov}(g^{(1)},(s),\ g^{(1)}(t)) = \sigma^2 e^{-\beta|s-t|} \quad (0 < \sigma^2 < \infty; 0 < \beta < \infty) \qquad (1.8)$$

of Orstein-Uhlenbeck kernel yielded sensible results in cases (a) and (b), as discussed by Leonard (1978), a paper which we will refer to as (L). The latter is a continuous version of the method for histograms discussed by Leonard (1973).

Note that the prior parameter σ^2 in (1.8) measures the closeness of $g^{(1)}$ to its prior estimate $\mu^{(1)}$ and β measures the smoothness of $g^{(1)}$ and g. These may be referred to as the "shrinkage" and "smoothing" parameters; we will later describe how they may be estimated empirically from the data.

The prior mean value function μ may be taken to represent the statistician's null hypothesis for μ. For example, in the biossay situation (d) he might take

$$\mu(t) = \alpha + \delta t \quad for \quad t \in (0,\infty)$$

representing a hypothesized logistic linear model. If any unknown parameters appear in the null hypothesis then they could be estimated by standard parametric techniques, under the hypothesized assumption that μ represents the true model.

For the Poisson process (b) and the time series example in (c) the extra, Gaussian, stage, to the distributional assumptions, could be taken as representing further time dependence in the sampling model. For example, in (b) the two stages of the distributional assumptions together give a broad representation of doubly stochastic Poisson processes (see Snyder (1975)) whilst in (e) we obtain a large class of possibly non-linear hierarchical time series models of the Kalman (1961) type.

We will propose two possible methods of estimation for $g(\cdot)$. The first gives a "posterior maximum likelihood estimate" of g as the solution of a typically non-linear Fredholm integral equation. The second method will just be described under the particular covariance structure in (1.8). It promises to be useful operationally since it also permits the empirical estimation of the prior covariance parameters, σ^2 and β. This involves approximating g by a linear combination of known basis functions,, projecting our function space prior onto an appropriate prior for the unknown parameters in the linear combination, and completing a Bayesian/empirical Bayesian analysis in finite dimensional space. The first procedure is summarized in the next section.

2. POSTERIOR ESTIMATION OF g.

By an extension to a result described on page 141 of the discussion of I it is possible to estimate g by the solution $\breve{g}$ to the integral equation

$$\breve{g}(t) = \mu(t) + \int_a^b K(s,t)\eta(\breve{g},s)ds \qquad t \in (a,b) \tag{2.1}$$

where the function $\{\eta(\breve{g},t);\ t \in (a,b)\}$ is chosen to ensure that

$$\left.\frac{\partial L(\breve{g} + \varepsilon u)}{\partial \varepsilon}\right|_{\varepsilon=0} = \int_a^b \eta(\breve{g},s)u(s)ds \tag{2.2}$$

is zero for all integrable functions $\{u(t);\ t \in (a,b)\}$.

In our five special cases this yields the following choices for $\eta(\breve{g},t)$:

$$\text{(a)} \qquad \phi_n(t) - n\exp\{\breve{g}(t)\}/\int e^{\breve{g}(s)}ds \tag{2.3}$$

$$\text{(b)} \qquad \phi_n(t) - \exp\{\breve{g}(t)\} \tag{2.4}$$

$$\text{(c)} \qquad \phi_r(t) - \sum_{i=1}^{n} I_{[t < y_i]}(t)\exp\{\breve{g}(t)\} \tag{2.5}$$

$$\text{(d)} \qquad \sum_{i=1}^{n} [y_i - \exp\{g(x_i)\}/(1 + \exp\{g(x_i\}))]\delta_{x_i}(t) \tag{2.6}$$

and

$$\text{(e)} \qquad \sum_{i=1}^{n} \frac{\partial \log b[y_i; g(t_i)]}{\partial g(t_i)} \delta_{t_i}(t) \tag{2.7}$$

where

$$\phi_n(t) = \sum_{i=1}^{n} \delta_{y_i}(t) \tag{2.8}$$

with $\delta_{y_i}(t)$ denoting the Dirac-delta function at $t = y_i$ and $I_{[t<y_i]}(t)$ the indicator function for $[t < y_i]$.

In the biossay example (d) we obtain from (2.1) the equation

$$\breve{g}(t) = \mu(t) + \sum_{i=1}^{n} K(x_i,t)[y_i - e^{\breve{g}(x_i)}/(1+e^{\breve{g}(x_i)})] \tag{2.9}$$

$$\text{for} \quad t \in (a,b).$$

In this case, and similarly in (e), the n+1 quantities $\breve{g}(t), \breve{g}(x_i), \cdots, \breve{g}(x_n)$ should be interpreted as the joint posterior modes of $g(t), g(x_1), \cdots, g(x_n)$. This is because equation (2.9) may be easily obtained by multiplying the likelihood functional exponentiating (1.5) by the (n+1)-

dimensional multivariate normal prior distribution of these quantities, and then maximizing the consequent posterior distribution.

Equation (2.9) may be solved computationally by firstly solving the n-dimensional non-linear system obtained by replacing t by each x_i in turn. Newton-Raphson will be adequate for solving this system for $\check{g}(x_1), \cdots, \check{g}(x_n)$. Then, in terms of these n quantities, (2.9) provides an explicit interpolation/extrapolation formula for $\check{g}(t)$, so that the complete continuum may be immediately estimated.

Note that the solution $\check{g}(t)$ to (2.9) will, for a smooth covariance kernel, possess similarly smooth regularity properties. For example, under the choice in (1.8) the equations become

$$\check{g}(t) = \mu(t) + \frac{\sigma^2}{\beta^2} \sum_{i=1}^{n} [\exp(-\beta(t-x_i)) + 2\beta(t-x_i)I_{[x_i<t]}(t)][y_i - \frac{e^{\check{g}(x_i)}}{1+e^{\check{g}(x_i)}}]$$

$$+ \text{ further terms not involving } t. \qquad (2.10)$$

Upon differentiating the expression on the right hand side it is easily checked that $\check{g}$ possesses a continuous second derivative. This is preferable to choosing the covariance kernel in (1.8) for g rather than $g^{(1)}$ since this leads to an estimate with a discontinuous first derivative.

Under very wide regularity conditions it is easy to show that the solutions for $\check{g}(x_1), \cdots, \check{g}(x_n)$ in (2.9) become strongly consistent as $n \to \infty$ for the corresponding true values $g(x_1), \cdots, g(x_n)$. We simply require the almost sure convergence of $m^{-1} \sum K(x_i, t) y_i$ to its sampling expectation i.e.

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} K(x_i, t) \frac{e^{g(x_i)}}{1+e^{g(x_i)}} \; . \qquad (2.11)$$

If this is true then the second term on the right hand side of (2.9) will become arbitrarily large, and the expression in (2.11) will therefore become equated with a similar expression but with the $g(x_i)$ replaced by the corresponding $\check{g}(x_i)$. Hence a consistent set of solutions will exist.

It is necessary to be more careful when interpreting our estimates in cases (a), (b), and (c), since the integral equations can no longer be reduced to an $n+1$ dimensional finite dimensional system. The best interpretation follows from a suggestion by Whittle on page 136 of the discussion of (L). He describes how our integral equations may be derived as the limit of a high-dimensional non-linear system obtained by discretization of the interval (a,b), and working with the high dimensional multivariate normal prior distribution discretizing our Gaussian prior on function space. In the discretized case the solutions of the non-linear system are the modes maximizing the posterior density. In the function space limit it is no longer meaningful to think in terms of modes since the posterior Radon-Nikodym derivative depends upon a choice of dominating measure. So the estimates may be referred to as "limiting posterior modes".

The integral equations in these three cases can alternatively be derived, for a fairly wide range of covariance kernels, via a direct optimization scheme on function space, involving prior and posterior "likelihoods" - see pages 117-8 of (L). Hence our estimates can also be interpreted as posterior maximum likelihood estimates.

Note that the regularity and consistency properties described above for case (d) also hold in the other five cases; in the first three cases the whole continuum of $\check{g}$ will be consistent for g.

3. SOME APPROXIMATE METHODS

In cases (a), (b), and (c) it is particularly tedious to solve the non-linear equations iteratively, and more difficult to find a procedure estimating any hyperparameters appearing in the covariance kernel. We therefore demonstrate how to obtain approximate solutions, under the special assumption in (1.8), but for all five cases (a), (b), (c), (d), and (e). These methods are particularly useful when n is large.

In (L) I show that, under this Orstein-Uhlenbeck covariance structure, the solution for $\tilde{g}$ to (2.1) is a maximizer of the functional

$$L_1(g) = L(g) + L_0(g) \tag{3.1}$$

where

$$\begin{aligned}-2L_0(g) &= \frac{1}{\beta\sigma^2} \int_a^b \{g^{(2)}(t) - \mu^{(2)}(t)\}^2 dt \\ &\quad + \frac{\beta}{\sigma^2} \int_a^b \{g^{(1)}(t) - \mu^{(1)}(t)\}^2 dt \\ &\quad + \frac{1}{\sigma^2} \{g^{(1)}(a) - \mu^{(1)}(a)\}^2 + \frac{1}{\sigma^2} \{g^{(1)}(b) - \mu^{(1)}(b)\}^2 \end{aligned} \tag{3.2}$$

is obtained via a Radon-Nikodym derivative for the prior distribution. Note that $L_1(g)$ in (3.1) plays the role of a penalized log-likelihood, since $-2L_0(g)$ may be interpreted as a roughness penalty. This term penalizes high first and second derivatives for the estimated g function.

Suppose now that we approximate $\tilde{g}$ by the linear functional

$$g^*(t) = \gamma^T \phi(t) \text{for } t \in (a,b) \tag{3.3}$$

where γ is a $p \times 1$ vector of unknown parameters, and $\phi(t)$ is a $p \times 1$ vector of known functions. For example, ϕ could consist of p orthogonal polynomials; alternately ϕ could contain the basis of B-splines for which g^* is a general cardinal spline with p knots. Note that this choice need not in practice be related to regularity conditions on the "true" function; since the true function is hypothetical enough for us to be able to fix our own regularity conditions.

Under the linear constraints in (3.3), maximization of $L_1(g)$ with respect to g is equivalent to maximization of $L_1(\utilde{\gamma})$ with respect to $\utilde{\gamma}$, where

$$L_1(\utilde{\gamma}) = L(\utilde{\gamma}) + L_0(\utilde{\gamma}) \tag{3.4}$$

where

$$-2L_0(\utilde{\gamma}) = (\utilde{\gamma} - \utilde{\mu})^T \utilde{R}(\utilde{\gamma} - \utilde{\mu}) \tag{3.5}$$

with

$$\utilde{R} = \frac{1}{\beta\sigma^2} \utilde{A}_2 + \frac{\beta}{\sigma^2} \utilde{A}_1 + \frac{1}{\sigma^2} \utilde{A}_0 \tag{3.6}$$

and

$$\utilde{R}\, \utilde{\mu} = \frac{1}{\beta\sigma^2} \utilde{d}_2 + \frac{\beta}{\sigma^2} \utilde{d}_1 + \frac{1}{\sigma^2} \utilde{d}_0 \tag{3.7}$$

where

$$\utilde{A}_2 = \int \utilde{\phi}^{(2)}(t)\utilde{\phi}^{(2)T}(t)dt$$

$$\utilde{A}_1 = \int \utilde{\phi}^{(1)}(t)\utilde{\phi}^{(1)T}(t)dt$$

$$\utilde{A}_0 = \utilde{\phi}^{(1)}(a)\utilde{\phi}^{(1)T}(a) + \utilde{\phi}^{(1)}(b)\utilde{\phi}^{(1)T}(b)$$

$$\utilde{d}_2 = \int \mu^{(2)}(t)\utilde{\phi}^{(2)}(t)dt$$

$$\utilde{d}_1 = \int \mu^{(1)}(t)\utilde{\phi}^{(1)}(t)dt$$

$$\utilde{d}_0 = \mu^{(1)}(a)\utilde{\phi}^{(1)}(a) + \mu^{(1)}(b)\utilde{\phi}^{(1)}(b).$$

The function in (3.4) will be minimized when $\utilde{\gamma} = \check{\utilde{\gamma}}$, where

$$\frac{\partial L(\check{\utilde{\gamma}})}{\partial \check{\utilde{\gamma}}} = \utilde{R}(\check{\utilde{\gamma}} - \utilde{\mu}). \tag{3.8}$$

In the five cases introduced in the last section, the derivative on the left hand side of (3.8) is given by

(a) $$\utilde{t} - n \int_a^t \utilde{\phi}(t)\exp\{\check{\utilde{\gamma}}^T \utilde{\phi}(t)\}dt \Big/ \int_a^b e^{\check{\utilde{\gamma}}^T \utilde{\phi}(t)} dt \tag{3.9}$$

where $\utilde{t} = \sum_{i=1}^{n} \utilde{\phi}(x_i)$

(b) $$\underset{\sim}{t} - \int_a^b \underset{\sim}{\phi}(t)\exp\{\breve{\gamma}^T\underset{\sim}{\phi}(t)\}dt \qquad (3.10)$$

$$\text{where } \underset{\sim}{t} = \sum_{i=1}^{n} \underset{\sim}{\phi}(x_i)$$

(c) $$\underset{\sim}{t} - \sum_{i=1}^{n} \int_a^{y_i} \underset{\sim}{\phi}(t)\exp\{\breve{\gamma}^T\underset{\sim}{\phi}(t)\}dt \qquad (3.11)$$

$$\text{where } \underset{\sim}{t} = \sum_{i=1}^{r} \underset{\sim}{\phi}(x_i)$$

(d) $$\underset{\sim}{t} - \sum_{i=1}^{n} \underset{\sim}{\phi}(x_i) \frac{e^{\breve{\gamma}^T\underset{\sim}{\phi}(x_i)}}{1+e^{\breve{\gamma}^T(x_i)}} \qquad (3.12)$$

$$\text{where } \underset{\sim}{t} = \sum_{i=1}^{n} y_i\underset{\sim}{\phi}(x_i)$$

and

(e) $$\sum_{i=1}^{n} \underset{\sim}{\phi}(t_i) \frac{\partial}{\partial u} b(y_i,u)\Big|_{u = \breve{\gamma}^T\underset{\sim}{\phi}(t)} \;. \qquad (3.13)$$

Note that, in the first four cases, there are p-dimensional sufficient statistics for $\underset{\sim}{\gamma}$, given the approximations. In the spline case, $\underset{\sim}{t}$ will consist of sample B-splines, and in the polynomial case it will contain sample moments. Therefore the choice of basis functions should perhaps depend upon which statistics are thought to be most relevant.

Equation (3.8) may in general be solved by Newton-Raphson, combined with an integration routine for cases (a), (b), and (c). The convergence will be faster than ordinary maximum likelihood for $\underset{\sim}{\gamma}$ owing to the normalizing effect on the Hessian of the $\underset{\sim}{R}$ matrix on the right hand side. Then the solution $\breve{g}$ to (2.1) may be approximated by $\breve{\gamma}^T\underset{\sim}{\phi}(t)$. A spline basis gives excellent error terms; p = 10 or 15 should be adequate; the choice of p can be juggled to check accuracy of the computations. This seems to give a computationally useful way of approximating the solution to our non-linear integral equations.

4. APPROXIMATE POSTERIOR PROBABILITIES

Owing to the simplicity of the quadratic form in (3.5) our approximate estimation procedure is equivalent to an exact Bayesian analysis for $\boldsymbol{\gamma}$ under the specified log likelihood $L(\boldsymbol{\gamma})$, and the multivariate normal prior

$$\boldsymbol{\gamma} \sim N(\boldsymbol{\mu}, \boldsymbol{R}^{-1}). \tag{4.1}$$

We therefore propose the mean vector and precision matrix in (3.6) and (3.7) as possessing reasonable structures for meaningfully representing prior information about the coefficients in a linear approximation. We are unfamiliar with other candidates for this, even in the straightforward polynomial regression situation.

Under this finite dimensional Bayesian analysis the exact posterior density of $\boldsymbol{\gamma}$ is

$$\pi(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \exp\{L(\boldsymbol{\gamma}) - \tfrac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu})^T \boldsymbol{R}(\boldsymbol{\gamma} - \boldsymbol{\mu})\} \tag{4.2}$$

and $\check{\boldsymbol{\gamma}}$ satisfying (3.8) is the exact posterior mode vector. The mode vector $\check{\boldsymbol{\gamma}}$ is well-known to be a first approximation to the posterior mean vector; refinements may if necessary be obtained via an Edgeworth expansion to the posterior density. Similarly the posterior covariance matrix of $\check{\boldsymbol{\gamma}}$ may be approximated by the exact posterior dispersion matrix $\boldsymbol{D}$ (the inverse of the Hessian in the Newton-Raphson iterations for solving (3.8)), where

$$\boldsymbol{D}^{-1} = \frac{-\partial^2 \log\pi(\check{\boldsymbol{\gamma}}|\boldsymbol{y})}{\partial(\check{\boldsymbol{\gamma}}\check{\boldsymbol{\gamma}}^T)} = \frac{-\partial^2 L(\check{\boldsymbol{\gamma}})}{\partial(\check{\boldsymbol{\gamma}}\check{\boldsymbol{\gamma}})^T} + \boldsymbol{R}. \tag{4.3}$$

This gives the following approximation to the posterior distribution, which could, if required, be replaced by exact expansions:

$$\boldsymbol{\gamma}|\boldsymbol{y} \sim N(\check{\boldsymbol{\gamma}}, \boldsymbol{D}). \tag{4.4}$$

Hence the posterior distribution of the g function is approximately Gaussian with mean value function

$$\check{g}(t) = \check{\boldsymbol{\gamma}}^T \boldsymbol{\phi}(t) \qquad (4.5)$$

and covariance kernel

$$\mathrm{cov}(\check{g}(s), \check{g}(t)) = \boldsymbol{\phi}^T(s) \boldsymbol{D} \boldsymbol{\phi}(t) \qquad (4.6)$$

yielding the possibility of very general posterior probability statements about g.

5. THE ESTIMATION OF σ^2 and β

It is possible to estimate σ^2 and β by approximating the values maximizing their "marginal log-likelihood"

$$\log p(\boldsymbol{y}|\sigma^2,\beta) = \log \int p(\boldsymbol{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}|\sigma^2,\beta)d\sigma^2 d\beta. \tag{5.1}$$

Note that, using similar notation, it is possible to calculate (5.1) approximately, using

$$\begin{aligned}\log p(\boldsymbol{y}|\sigma^2,\beta) &= \log p(\boldsymbol{y}|\tilde{\boldsymbol{\gamma}}) + \log \pi(\tilde{\boldsymbol{\gamma}}|\sigma^2,\beta) - \log \pi(\tilde{\boldsymbol{\gamma}}|\boldsymbol{y},\sigma^2,\beta) \\ &\approx L(\tilde{\boldsymbol{\gamma}}) - \tfrac{1}{2}\log|\boldsymbol{D}| - \tfrac{1}{2}\log|\boldsymbol{R}| + \tfrac{1}{2}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\mu})^T\boldsymbol{R}(\boldsymbol{\gamma} - \boldsymbol{\mu}).\end{aligned} \tag{5.2}$$

This approximation is based upon a backwards application of Bayes theorem together with the normal approximation to the posterior density of $\boldsymbol{\gamma}$. Hence it is possible to calculate an approximation to the whole marginal log-likelihood of σ^2 and β.

In order to obtain point estimates for σ^2 and β we refer to the following lemma, which could be viewed as a special case of the EM algorithm; see Dempster, Laird, and Rubin (1977).

<u>Lemma:</u> Suppose that the sampling distribution of a vector of observations $\boldsymbol{y}$ depends upon p unknown parameters $\gamma_1,\cdots,\gamma_p$ and that the prior distribution of $\gamma_1,\cdots,\gamma_p$ depends upon $r < p$ prior parameters $\beta_1,\cdots,\beta_r$. Suppose further that with $q > r$ the prior density of $\gamma_1,\cdots,\gamma_p$ belongs to the q-parameter exponential family

$$\pi(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \exp\{\sum_{j=1}^{q} v_j(\boldsymbol{\beta})t_j(\boldsymbol{\gamma}) + a(\boldsymbol{\beta})\} \tag{5.3}$$

where a and the v_j do not depend upon $\boldsymbol{\gamma}$ and the t_j do not depend upon $\boldsymbol{\beta}$. Then, if the marginal likelihood of $\boldsymbol{\beta}$ is differentiable, it is maximized by values satisfying the equation

$$\sum_{j=1}^{q} \frac{\partial v_j(\boldsymbol{\beta})}{\partial\beta_k}\left[E[t_j(\boldsymbol{\gamma})|\boldsymbol{\beta}] - E[t_j(\boldsymbol{\gamma})|\boldsymbol{\beta},\boldsymbol{y})\right] = 0 \qquad (k = 1,\cdots,p). \tag{5.3}$$

The proof of this lemma is straightforward upon differentiating

$$\log p(\boldsymbol{y}|\boldsymbol{\beta}) = \log \int p(\boldsymbol{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}|\boldsymbol{\beta})d\boldsymbol{\gamma}$$

with respect to $\boldsymbol{\beta}$.

To apply the lemma note from (3.2) and (3.3) that the prior may be expressed in the required form with $r=2$, $y=3$, $\beta_1 = \sigma^{-2}$, $\beta_2 = \beta$, $v_1 = \beta_1\beta_2$, $v_2 = \beta_1\beta_2^{-1}$, and $v_3 = \beta_1$. After some manipulations the equations reduce, under our approximation to the posterior distribution, to

$$\sigma^2 = p^{-1}(\breve{\underset{\sim}{Y}} - \underset{\sim}{\mu})^T(\breve{\underset{\sim}{Y}} - \underset{\sim}{\mu}) + p^{-1}\ \mathrm{trace}(\underset{\sim}{R}\underset{\sim}{D}) \qquad (5.3)$$

and

$$\beta^2 = \frac{(\breve{\underset{\sim}{Y}} - \underset{\sim}{\mu}_2)^T\ \underset{\sim}{A}_2(\breve{\underset{\sim}{Y}} - \underset{\sim}{\mu}_2) + \mathrm{trace}(\underset{\sim}{A}_2\underset{\sim}{D}) - (\underset{\sim}{\mu} - \underset{\sim}{\mu}_2)^T\underset{\sim}{A}_2(\underset{\sim}{\mu} - \underset{\sim}{\mu}_2)^T - \mathrm{trace}(\underset{\sim}{R}_2^{-1}\underset{\sim}{A}_2)}{(\breve{\underset{\sim}{Y}} - \underset{\sim}{\mu}_1)^T\underset{\sim}{A}_1(\breve{\underset{\sim}{Y}} - \underset{\sim}{\mu}_1) + \mathrm{trace}(\underset{\sim}{A}_1\underset{\sim}{D}) - (\underset{\sim}{\mu} - \underset{\sim}{\mu}_1)^T\underset{\sim}{A}_1(\underset{\sim}{\mu} - \underset{\sim}{\mu}_1) - \mathrm{trace}(\underset{\sim}{R}_1^{-1}\underset{\sim}{A}_1)} \qquad (5.4)$$

where

$$\underset{\sim}{\mu}_1 = \underset{\sim}{A}_1^{-1}d_1$$

and

$$\underset{\sim}{\mu}_2 = \underset{\sim}{A}_2^{-1}d_2$$

with the $\underset{\sim}{A}$'s and $\underset{\sim}{d}$'s defined in (3.7). Equations (5.3) and (5.4) may be solved by cyclic substitutions in conjection with (3.8) and (4.3).

6. FILTERING A NORMAL MEAN VALUE FUNCTION OR REGRESSION FUNCTION.

We now consider a very special case of the optimization problem in (3.1) and (3.2) which will be relevant when the observations $y_1,\cdots,y_m$ are independent and normally distributed, given their respective means $g(t_1),\cdots,$ $g(t_m)$ and common variance τ^2. Consider the optimization with respect to g of

$$L_1(g) = \tau^{-2} \sum_{i=1}^{n} (y_i - g(t_i)^2) + \phi^{-1} \int_a^b [g^{(2)}(t)]^2 dt. \qquad (6.1)$$

The first term on the right hand side relates to the sampling log-likelihood under our normal theory assumptions. The second term may be obtained by letting $\sigma^2 \to \infty$ and $\beta\sigma^2 \to \phi$ in our Orstein-Uhlenbeck process (3.2), and represents an integrated Wiener process.

The optimizer $\check{g}$ of (6.1) is well-known (e.g. Wahba, 1975,76) to be a cubic spline with n knots. We now consider the estimation of τ^2 and ϕ^{-1} and consider linear combinations of the form described in (3.3) where $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}(t)$ are p-dimensional. Note that when $p = n$ this assumption is exact under the appropriate choices of basis functions for $\check{\boldsymbol{\phi}}(t)$ (which leads to a general B-spline for g with n knots and no constraints on the coefficients). When p is fixed to be around 10 or 15 we have approximations which will be particularly useful when n is large. However, the theory described below relates to both of these exact and approximate situations.

Under the linear assumption in (3.3), (6.1) may be replaced by

$$L_1(\boldsymbol{\gamma}) = \tau^{-2}(\boldsymbol{\gamma} - \boldsymbol{\ell})^T \boldsymbol{P}(\boldsymbol{\gamma} - \boldsymbol{\ell}) + \gamma^{-1}\boldsymbol{\gamma}^T \boldsymbol{A}_2 \boldsymbol{\gamma} \qquad (6.3)$$

where $\boldsymbol{A}_2$ is given after (3.7),

$$\boldsymbol{\ell} = \boldsymbol{P}^{-1}\boldsymbol{b} \qquad (6.4)$$

$$\boldsymbol{P} = \sum_{i=1}^{m} \boldsymbol{\phi}(t_i)\boldsymbol{\phi}^T(t_i) \qquad (6.5)$$

and

$$\mathbf{b} = \sum_{i=1}^{m} y_i \boldsymbol{\phi}(t_i). \qquad (6.6)$$

This reduces the optimization problem to the classical Bayesian problem of estimating a regression vector $\boldsymbol{\gamma}$ under a multivariate normal $N(\mathbf{0},\sigma^2\mathbf{A}_2^{-1})$ prior, giving the standard result

$$\boldsymbol{\gamma}|\mathbf{y} \sim N(\tilde{\boldsymbol{\gamma}},\mathbf{D}) \qquad (6.7)$$

where

$$\tilde{\boldsymbol{\gamma}} = \tau^{-2}(\tau^{-2}\mathbf{B} + \phi^{-1}\mathbf{A}_2)^{-1}\mathbf{b} \qquad (6.8)$$

$$= (\mathbf{B} + \lambda\mathbf{A}_2)^{-1}\mathbf{b}$$

with

$$\lambda = \tau^2/\phi \qquad (6.9)$$

and

$$\mathbf{D}^{-1} = \tau^{-2}(\mathbf{B} + \lambda\mathbf{A}_2). \qquad (6.10)$$

The parameters λ and τ^2 may now be estimated by any one of the following three procedures

(a) Application of the EM algorithm.

By a direct application of the EM algorithm described by Dempster et. al. it is possible to show that the marginal maximum likelihood estimates of τ^2 and ϕ satisfy the equations

$$\tau^2 = m^{-1}\sum_{i=1}^{m}(y_i - \tilde{\boldsymbol{\gamma}}^T\boldsymbol{\phi}(t_i))^2 + m^{-1}\mathrm{trace}(\mathbf{D}\mathbf{B}) \qquad (6.11)$$

and

$$\phi = p^{-1}\tilde{\boldsymbol{\gamma}}^T\mathbf{A}_2\tilde{\boldsymbol{\gamma}} + p^{-1}\ \mathrm{trace}(\mathbf{D}). \qquad (6.12)$$

Equations (6.11) and (6.12) may be solved by cyclic substitutions. Evaluate (6.8) and (6.10) for trial values of τ^2 and ϕ^2, substitute for $\tilde{\boldsymbol{\gamma}}$ and $\mathbf{D}$ on the right hand sides of (6.11) and (6.12), obtain new values for τ^2 and ϕ on the left hand side and repeat the procedure until convergence. The latter is guaranteed under general results governing the EM algorithm.

Setting $\lambda = \tau^2/\phi$ the above procedure will achieve the maximum of the marginal likelihood obtained by noting that the marginal distribution of $\underset{\sim}{\ell}$, given τ^2, and λ is multivariate normal with zero mean vector and covariance matrix $\tau^2(\underset{\sim}{P}^{-1} + \lambda^{-1}\underset{\sim}{A}_2^{-1})$. The marginal likelihood is

$$\ell(\tau^2,\lambda|y) = p(\underset{\sim}{y}|\tau^2,\lambda)$$

$$\propto (\tau^2)^{-\frac{1}{2}n}\lambda^{\frac{1}{2}p} \exp\{-\tfrac{1}{2}\tau^{-2}\underset{\sim}{\lambda}\underset{\sim}{\ell}\underset{\sim}{A}_2(\underset{\sim}{A}_2 + \underset{\sim}{\lambda}\underset{\sim}{P})^{-1}\underset{\sim}{P}\underset{\sim}{\ell}\} \qquad (6.13)$$

for which more standard optimization procedures might prove tedious.

(b) <u>Bayesian Methods</u>

Suppose that $\nu\omega/\tau^2$ possesses a prior distribution which is chi-squared with ν degrees of freedom and which is independent of λ. Then ω^{-1} is the prior mean of τ^{-2} and ν is the prior 'sample size'. In ignorance situations a small but non-zero value e.g. $\nu = 1$ should be chosen for ν. For general ν we find from (6.13) that the marginal posterior density of λ is

$$\pi(\lambda|\underset{\sim}{y}) \propto \pi(\lambda)\lambda^{\frac{1}{2}p}[\nu\omega + \lambda\underset{\sim}{\ell}\underset{\sim}{A}_2(\underset{\sim}{A}_2 + \lambda\underset{\sim}{P})^{-1}P\underset{\sim}{\ell}\underset{\sim}{\iota}]^{-\frac{1}{2}(\nu+n)} \quad (0 < \lambda < \infty) \qquad (6.14)$$

where $\pi(\lambda)$ may be set proportional to unity in ignorance situations, with any proper prior distribution permissible in the presence of prior knowledge about λ. Note that Bayes estimates for λ may be easily obtained from (6.14). For example, the posterior mean may be obtained via the obvious one-dimensional integration. Furthermore the posterior mean of τ^{-2} is

$$E(\tau^{-2}|\underset{\sim}{y}) = \int E(\tau^{-2}|\lambda,\underset{\sim}{y})\pi(\lambda|\underset{\sim}{y})d\lambda \qquad (6.15)$$

where

$$E(\tau^{-2}|\lambda,\underset{\sim}{y}) = (\nu + n)/[\nu\omega + \lambda\ell\underset{\sim}{A}_2(\underset{\sim}{A}_2 + \lambda\underset{\sim}{P})^{-1}\underset{\sim}{P}\underset{\sim}{\ell}] \qquad (6.16)$$

and the unconditional posterior mean vector of $\underset{\sim}{y}$ is

$$E(\underset{\sim}{y}|y) = \int E(\underset{\sim}{y}|\lambda,\underset{\sim}{y})\pi(\lambda|y)d\lambda \qquad (6.17)$$

where the first contribution to the integrand is given in (6.8).

These results may be generalized, as required, to give the unconditional covariance matrix and posterior distribution of $\underset{\sim}{\gamma}$ and the unconditional mean value, covariance kernel, together with unconditional posterior probabilities, for the function $g(t) = \underset{\sim}{\gamma}^T\phi(t)$.

(c) Cross-Validation.

Wahba (1976) and others estimate λ by an empirical cross-validation procedure which does not relate to the likelihood function in (6.13). We imagine that her procedure will prove appealing when the statistician is not clear about his particular choice of prior functional in (6.1).

REFERENCES

Kalman, R. E. A new approach to linear filtering and prediction problems. Trans. SME Ser D. 82, 35-45.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood for incomplete data (with Discussion) J. R. Statist. Soc. B 39, 1-48.

Leonard, T. (1973). A Bayesian method for histograms. Biometrika 60, 297-208.

Leonard, T. (1978) Density Estimation, Stochastic Processes, and Prior Information (with Discussion), (J. Roy. Statist. Soc. B 40, 113-146.

Snyder, D. L. (1975) Random Point Processes New York. J. Wiley and Sons.

Wahba, G. (1975) Interpolating spline methods for density estimation, Ann. Statist., 3, 30-48.

Wahba, G. (1976). Optimal smoothing of density estimates. Department of Statistics, University of Wisconsin, Madison Technical Report No. 469.

TL/db

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2339	2. GOVT ACCESSION NO. AD-A114573	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AN EMPIRICAL BAYESIAN APPROACH TO THE SMOOTH ESTIMATION OF UNKNOWN FUNCTIONS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Tom Leonard		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE February 1982 13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Empirical Bayes; function space; density estimation; non-homogeneous Poisson process; survival data; hazard function; biossay; logit; time series; non-linear filtering; Kalman; Gaussian process; Orstein-Uhlenbeck; covariance kernel; EM algorithm.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A Bayesian procedure is described for smoothly estimating unknown functions, given a finite set of observations. It is assumed that a suitable transformation of the function can be taken to possess a Gaussian prior distribution across function space. The five special cases (a) estimation of a logistic density transform, (b) the log intensity function of a non-homogeneous Poisson process, (c) the log hazard function for survival data, (d) the logit function in biossay, and (e) the mean value function in a possibly non-linear time series of the Kalman type or equivalently a —>

20. Abstract (continued)

regression function for possibly non-normal observations, are considered, and in each case a non-linear Fredholm equation is described for the posterior estimate. In two cases ~~(d) and (e)~~ this reduces to a finite dimensional system. In all five cases an approximate procedure is developed which is particularly useful when the sample size is large. This approximates the function space prior by a multivariate normal prior on the coefficients in a linear approximation, and then proceeds by conventional Bayesian techniques. In cases where the prior covariance kernel is assumed to posses a particular parametrized form (e.g. from the Orstein-Ulenbeck process) the approximations enable us to estimate the prior parameters appearing in the kernel, empirically from the data, via a lemma relating to the EM algorithm. Finally, in a very special case, involving normal observations and an integrated Wiener prior, some fresh Bayesian and empirical Bayesian results are developed.